# Обучение машинного перевода без параллельных текстов*

***Ярошенко А. М. Бахтеев*[1] *О. Ю. Стрижов*[2] *В. В.***
[1]Московский физико-технический институт
[2]Вычислительный центр им. А. А. Дородницына ФИЦ ИУ РАН

В данной работе исследуется задача машинного перевода между двумя языками. Для решения часто используются параллельные предложения, то есть совпадающие по смыслу фразы на двух языках. В работе рассматривается альтернативная модель, не требующая большого количества параллельных предложений. Она использует нейронную сеть типа Seq2Seq, имеющую скрытое пространство. [Тут добавится что-то от меня]. Для проверки качества модели проводится вычислительный эксперимент по переводу предложений между близкими языками, такими как русский и украинский.

**Ключевые слова**: *нейронные сети, машинный перевод, автокодировщики.*

## Введение

В зависимости от специфики пары языков выделяют несколько подходов к машинному переводу. При наличии достаточного числа параллельных предложений(порядка миллиона) использование глубоких нейронных сетей привело к получению хороших результатов [1],[2].

Но для многих пар языков нет достаточной базы примеров. Одним из подходов на основе параллельных предложений является пополнение обучающей выборки переводами с предыдущих итераций работы нейронной сети [3].

Ниже представлено решение задачи машинного перевода при отсутствии достаточного количества параллельных предложений [4], [5], [6]. В модели используются 2 типа автокодировщиков: рекуррентные нейронные сети [7], [8], которые реализуют перевод слов в скрытое пространство, и сеть-дискриминатор, определяющая по векторному представлению язык исходного предложения. Сети-энкодеры оптимизируются так, чтобы представление одного и того же предложения на разных языках совпадало в скрытом пространстве, то есть, чтобы дискриминатору было сложнее определить язык, к которому относится вектор. Обучение состоит из двух фаз. На первой оптимизируется работа дискриминатора: предложение кодируется с добавлением шума [9] и подаётся на вход и происходит перераспределение параметров. На второй стадии происходит перераспределение парамтеров уже у сетей-кодировщиков. После проведение этих шагов вычисляется значение функции потерь.

Такой подход был использован в [10] для пары языков французский-английский. В данной работе будет проведен схожий эксперимент для перевода с русского на украинский. Качество переводчика в работе оценивается с помощью метрики BLEU [11].

## Постановка задачи

В данной задаче в качестве обучающей выборки используются несопоставленные друг другу предложения на обоих языках $D^{src} = [s_1^{src}, ..., s_{m_{src}}^{src}]$, $D^{tgt} = [s_1^{tgt}, ..., s_{m_{tgt}}^{tgt}]$, по которым

Научный руководитель: Стрижов В. В. Авторы: А.В. Грабовой, О.Ю. Бахтеев, В.В. Стрижов, Eric Gaussier, координатор Малиновский Г. С. Консультант: Бахтеев О. Ю.

нужно предоставить перевод на другой язык. Также предоставлен корпус параллельных предложений $D^{valid} = \{(s_1^{src}, s_1^{tgt}), ..., (s_{m_{valid}}^{src}, s_{m_{valid}}^{tgt})\}$ для проверки качества перевода.

Предлагается решение в виде модели $\mathbf{M}$ из двух рекуррентых нейронных сетей для реализации декодера $\mathbf{g}$ и энкодера $\mathbf{f}$ и из двуслойного персептрона $\mathbf{discr}$ в качестве дискриминатора.

В качестве метрики между словами используется NLL. Метрикой качества модели является среднее значение BLEU на валидационной выборке:

$$\frac{1}{m_{valid}} \sum_{i=1}^{m_{\text{valid}}} BLEU(M(s_i^{src}), s_i^{tgt})$$

Так как у нас нет достаточно большого корпуса из параллельных предложений, мы будем использовать следующую схему построения модели. Используется два словаря $V^{src}$ и $V^{tgt}$, которые сопосталяют словам из обоих языков численные векторы. Предварительно по выборке строится нулевое приближение модели - пословный перевод. Для этого каждому слову $\mathbf{x}$ ставится в соответствие его ближайший по косинусной метрике сосед $\mathbf{y} = argmin_z \rho(x, z)$ из векторного представления MUSE.

При последующих итерациях энкодер будет переводить исходное предложение в скрытое пространство $Z$, общее для обоих языков, где дискриминатор будет по вектору определять, какому языку он принадлежал. Параметры оптимизируются так, чтобы дискриминатор по представлению предложения из языка source определял его как target, то есть чтобы латентные пространства языков были достаточно близки и усложнялась работа дискриминатора. Декодер же преобразует этот вектор из $Z$ в матрицу вероятностей, где для каждого слова $\mathbf{x}$ будет определен вектор вероятностей $\boldsymbol{p(x)}$ нехождения на этой позиции того или иного слова из словаря $V^{tgt}$. Параметры у энкодеров и декодеров для обоих языков являются общими, для этих моделей отличается только генерация матрицы вероятностей.

Помимо этого для избежания переобучения к каждому предложению добавляется шум $\sigma$: некоторые слова удаляются, а остальные переставляются, но так чтобы результирующая позиция отличалась от исходной не более, чем на фиксированную константу k.

Для реализации этого метода определим функционалы, которые будут минимизироваться. Во-первых, это ошибка восстановления зашумленного предложения в исходный язык, которая считается между входным предложением $\mathbf{s^{src}}$ и его образом $g^{src}(f^{src}(\mathbf{s^{src}}))$. На этом шаге оптимизации будет минимизироваться следующая функция:

$$L_{AE} = \sum_{i=1}^{|s^{src}|} NLL(g^{src}(f^{src}(\sigma(\mathbf{s^{src}})))[i], \mathbf{s^{src}}[i])$$

Далее рассматривается ошибка перевода уже на другой язык:

$$L_{TR} = \sum_{i=1}^{|s^{src}|} NLL(g^{tgt}(f^{src}(\sigma(\mathbf{s^{src}})))[i], \mathbf{s^{src}}[i])$$

И последний этап - оптимизация дискримантора для схожести латентных представлений одного и того же предложения, закодированного с разных языков:

$$L_{ADV} = \log \mathbb{P} discr(f^{src}(s^{src})) = src$$

Таким образом, имеем задачу оптимизации:

$$L = a * L_{AE} + b * L_{TR} + c * L_{ADV} \longrightarrow min$$

где a,b,c калибруемые гиперпараметры.

## Литература


[1] Bilingual word embeddings for phrase-based machine translation / Will Y Zou, Richard Socher, Daniel Cer [и др.] // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. 2013. С. 1393–1398.

[2] On the properties of neural machine translation: Encoder-decoder approaches / Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau [и др.] // arXiv preprint arXiv:1409.1259. 2014.

[3] Bertoldi Nicola, Federico Marcello. Domain adaptation for statistical machine translation with monolingual resources // Proceedings of the fourth workshop on statistical machine translation / Association for Computational Linguistics. 2009. С. 182–189.

[4] Google's neural machine translation system: Bridging the gap between human and machine translation / Yonghui Wu, Mike Schuster, Zhifeng Chen [и др.] // arXiv preprint arXiv:1609.08144. 2016.

[5] Sutskever Ilya, Vinyals Oriol, Le Quoc V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. 2014. С. 3104–3112.

[6] Bahdanau Dzmitry, Cho Kyunghyun, Bengio Yoshua. Neural machine translation by jointly learning to align and translate // arXiv preprint arXiv:1409.0473. 2014.

[7] Gers Felix A, Schmidhuber Jürgen, Cummins Fred. Learning to forget: Continual prediction with LSTM. 1999.

[8] Graves Alex, Schmidhuber Jurgen. Framewise phoneme classification with bidirectional LSTM and other neural network architectures // Neural Networks. 2005. Т. 18, № 5-6. С. 602–610.

[9] Kim Yunsu, Ney Jiahui Geng Hermann. Improving Unsupervised Word-by-Word Translation with Language Model and Denoising Autoencoder. 2018.

[10] Lample Guillaume, Denoyer Ludovic, Ranzato Marc'Aurelio. Unsupervised machine translation using monolingual corpora only // arXiv preprint arXiv:1711.00043. 2017.

[11] BLEU: a method for automatic evaluation of machine translation / Kishore Papineni, Salim Roukos, Todd Ward [и др.] // Proceedings of the 40th annual meeting on association for computational linguistics / Association for Computational Linguistics. 2002. С. 311–318.

[12] Word translation without parallel data / Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato [и др.] // arXiv preprint arXiv:1710.04087. 2017.